

From Predictions to Data-Driven Decisions Using Machine Learning

Nathan Kallus
Massachusetts Institute of Technology
77 Massachusetts Ave E40-149
Cambridge, MA 02139
kallus@mit.edu

ABSTRACT

Predictive analyses taking advantage of the recent explosion in the availability and accessibility of data have been made possible through flexible machine learning methodologies that are often well-suited to the variety and velocity of today's data collection. This can be witnessed in recent works studying the predictive power of social media data and in the transformation of business practices around data. It is not clear, however, how to go from expected-value predictions based on predictive observations to decisions that yield high profits and carry low risk. As classical problems of portfolio allocation and inventory management show, decisions based on mean-field analysis are suboptimal and high in risk. In this paper we endeavor to refit existing machine learning predictive methodology and theory to the purpose of prescribing optimal decisions based directly on data and predictive observations. We study the convergence as more data becomes available of such methods to the omniscient optimal decision, that which exploits these predictive observations to their fullest extent by using the unknown distribution of parameters. Incredibly, the data-driven prescriptions developed converge to the omniscient optimum for almost all realizations of data and for almost any given predictive observation and even when data is not IID, which is generally the case in practice. We consider an example of portfolio allocation to illustrate the power of these methods.

Keywords

Predictive analytics, Prescriptive analytics, Data-driven decision making, Machine learning

1. INTRODUCTION

The availability of machine-readable data has swelled in recent years. En-masse data collection is now a standard business practice. Online retailers record daily demands for thousands of products with ease and are able to track purchase behavior of individuals and tie it how they viewed

and clicked through their website and to the product reviews they post. Even traditional retailers and suppliers are able to observe order patterns across their multinational supply chains. At the same time more of our news is widely published online, enabling automated analysis of current situations with appropriate natural language processing. User-generated data is on the rise in Twitter posts and Facebook status updates and provides a peak into consumer opinion and the future of crowd behavior (see [24]). In 2012 about 2.5 exabytes of data were created each day and this number increases by some 25% each year, with Walmart consumer transactions alone making up approximately 0.01% of this (see [28]). The success of the Open Data movement has improved the accessibility of data, making climate, market, and governmental data and more available in one-stop shops and in standardized formats.

It is no wonder that machine learning and data mining has grown in importance in facilitating descriptive analyses (e.g. clustering) as well as predictive analyses (e.g. regression and classification) of such data leading to valuable insights. Such methodologies are also often well-suited to the volume and variety of today's data collection. This explosion of data coupled with methodological advances has enabled applications in business that predict consumer demand on Black Friday based on cell-phone location data [28] and applications in public health that track latent processes of disease spread based on online web-search queries [11]. Such queries have also been used to describe consumer behavior, most notably in [13] and [19], and to predict movements in the stock market in [14]. Online chatter on blogs and social media have been shown to predict movie earnings in [2] and Amazon book sales in [22] using IBM's WebFountain [21].

An important question, especially from a business's point of view, is how to go from predictions to decisions. The answer is not obvious. Consider a classical example in investment planning. A unit budget is to be invested in d securities, where security i will experience a yet-unknown return of Y_i percent over the investment period and z_i represents the fraction invested in i ($z_i \geq 0$, $\sum_i z_i = 1$). The returns on a portfolio z will then be $z^T Y$. In a predictive analysis, some realization of observable features $X = x$, such as recent returns, social media sentiment toward the underlying companies and their products, news coverage of these companies or the market at large, month of the year, analyst ratings, etc, would be used to construct a good predictor for Y . For example, we mentioned the use of social media data to predict demand at the box office [2], a quantity traded on platforms such as Intrade.com. A good predictor $\hat{y}(x)$ in

the predictive analytics sense is generally one that best approximates the conditional expectation function (a.k.a. regression function) $\mathbb{E}[Y|X = x]$, the average future value of Y in the particular cases when $X = x$. The procedure of constructing such a functional estimate based on historical data is known as regression analysis and is the most common type of predictive analysis of continuous quantities. In choosing a portfolio one is usually interested either in maximizing the expected value of a concave utility function in profits $\mathbb{E}[u(z^T Y)|X = x]$ or in minimizing a risk measure in the losses, such as conditional value at risk (see (2.2)). However, replacing Y by its conditional expectation in either case will result in a portfolio that invests the whole budget in a single security, which is generally suboptimal with respect to either of these objectives if u is nonlinear. Augmenting this with a data-driven matrix-functional estimate of conditional covariance $\text{Cov}(Y|X = x)$ (a non-trivial task) and constructing a conditional Markowitz portfolio that weighs expectation against variance could generally yield reasonable results but may not make full use of the available data and may not converge to the optimal policy had we known the distribution of Y conditioned on $X = x$ for any x .

Another example is in the classical model of the news vendor problem in inventory management augmented with predictive observations (such as recent product demands, social media sentiment toward the vendor and its product, negative news coverage of the vendor and positive news coverage of competitors, day of the week, month of the year, human expert predictions of demand, weather forecasts, etc) to forecast demand. Filling inventory to exactly match expected demand is suboptimal when the profits of selling outweigh the losses of overstocking. Another example is the general framework of two-stage problems augmented with predictive observations. An initial decision is made today informed by these and when future parameters realize we have a second chance to take a recourse that is limited by the initial decision. An agricultural example with uncertain crop yields and other examples of such two-stage problems in the classic stochastic optimization framework without predictive observations can be found in [7].

In this paper we endeavor to develop approaches that refit existing machine learning predictive methodology and theory to the purpose of prescribing optimal decisions based directly on data. We identify two main types of data-driven predictive-prescriptive methods. In similar ways to how flexible non-parametric machine learning methods can be used either for classification or for regression (e.g. k NN, decision trees, random forests, and boosting), we show how they can also be tailored for a third purpose: optimal decision-making under predictive observations. We call such methods conditional-distribution-based prescriptions. We study the convergence as more data becomes available of these data-driven prescriptions to the omniscient optimal decision—that which exploits these predictive observations to the fullest extent by using the unknown conditional distribution of parameters. Because the variety and velocity of modern data collection means that samples are usually never IID in nearly any practical application, we also study such convergence for data drawn from mixing processes such as ARMA, GARCH, and Markov chains, which model evolving systems like a stock market or a social network. Incredibly, the data-driven predictive prescriptions developed converge to the omniscient optimum for almost all realizations of data (al-

most surely) and for almost any given new observation $X = x$ (almost everywhere) even without IID data and often without assumptions on unknown distributions.

Another type of predictive prescription we identify is to optimally choose a decision rule that takes a functional form from within a family of possible ones so to minimize marginal empirical risk. We call these rule-based prescriptions. Here we borrow from the traditional machine learning theory of probability-approximately-correct (PAC)-learnability [40] in order to characterize out-of-sample performance in terms of costs or profits with appropriate generalizations and extensions of functional complexity notions to decision-valued rules. Again, we study the ramifications of non-IID data.

In data-poorer times, the application of mathematical optimization has largely relied on stylistic modeling of distributions with only marginal deference to data. This is especially true of predictive data although it can significantly improve performance by indicating what the future may hold. Nonetheless the theory and methods of optimization have utterly transformed entire industries: airlines, advertising, retail, finance, and more. Tapping the power of raw data, which is becoming so plentiful, in quantitative decision-making could trigger a second such revolution. We hope that the efforts presented here will be just one small step toward that.

2. CONDITIONAL-DISTRIBUTION-BASED PRESCRIPTIONS

We consider the following general set-up in this paper. We must make a decision $z \in \mathcal{Z} \subset \mathbb{R}^d$ today that carries a future cost of $c(z; Y)$, which depends on the value of unknown parameters Y taking values in $\mathcal{Y} \subset \mathbb{R}^{m_Y}$ that only materialize in the future. We have a synchronous observation of a random variable X taking values in $\mathcal{X} \subset \mathbb{R}^{m_X}$ that may help us predict Y and therefore help us choose z . We denote by μ the joint measure of X, Y , by $\mu_{Y|x}$ the conditional measure of Y given $X = x$, and by μ_X and μ_Y the marginal measure of X and Y , respectively. We will either be concerned with minimizing expected costs (where the cost function may incorporate disutility structure) or the conditional value at risk (CVaR), a popular risk measure, especially in financial applications. We assume throughout that $c(z; y)$ is μ_Y -integrable for every $z \in \mathcal{Z}$, that is, every feasible control z has a well-defined average future cost.

In the case of expected cost minimization, the optimal procedure after observing the present value x of X , had we known $\mu_{Y|x}$, would be the minimization problem

$$\min_{z \in \mathcal{Z}} \{C(z|x) := \mathbb{E}[c(z; Y)|X = x]\}. \quad (2.1)$$

The CVaR at level α of a random loss L with quantile function F_L^{-1} is the expectation above the $(1 - \alpha)$ -quantile:

$$\begin{aligned} \text{CVaR}_\alpha(L) &:= \mathbb{E}[L \mid L \geq F_L^{-1}(1 - \alpha)] \\ &= \inf_{\beta \in \mathbb{R}} \mathbb{E}\left[r_\alpha(L, \beta) := \beta + \frac{1}{\alpha}(L - \beta)_+\right] \end{aligned} \quad (2.2)$$

where the latter equivalent definition is due to Rockafellar and Uryasev [34] and $(v)_+ = \max\{v, 0\}$. Then, in the case of CVaR minimization, the optimal procedure after observing the present value x of X , had we known $\mu_{Y|x}$, would be the

minimization problem

$$\min_{z \in \mathcal{Z}, \beta \in \mathbb{R}} \left\{ R_\alpha(z, \beta|x) := \mathbb{E} \left[\beta + \frac{1}{\alpha} (c(z; Y) - \beta)_+ | X = x \right] \right\}. \quad (2.3)$$

Note that the integrality of $c(z; y)$ immediately implies the integrability of $r_\alpha(c(z; y), \beta)$ for each $z \in \mathcal{Z}, \beta \in \mathbb{R}$.

But we do not know $\mu_{Y|x}$. Instead we assume that we have a sample of previous observations of pairs of X, Y that help us learn the relationship between the two:

$$S_n = \{(x^1, y^1), \dots, (x^n, y^n)\}.$$

At times we will make assumptions on the generation of S_n (e.g. IID, mixing) to prove results. In the approach we study in this section we use this sample to construct a conditional-distribution estimator $\hat{\mu}_{Y|x,n}$ and plug it in place of $\mu_{Y|x}$ in (2.1) and (2.3) in order to choose our control z . The estimators we will use will always take the form of reweighting the existing sample of y 's based on the observation x :

$$\hat{\mu}_{Y|x,n} = \sum_{i=1}^n w_n^i(x) \delta_{y^i} \quad \text{for some } w_n^i(x) \geq 0, \sum_{i=1}^n w_n^i(x) = 1 \quad (2.4)$$

where δ_{y^i} denotes the Dirac measure at y^i . For example, the assumption that X is independent of Y would lead to the weights $w_n^i(x) = \frac{1}{n}$ and in turn to the standard sample average approximation (SAA) of stochastic programming (see [36, 37, 25]). Of course, if they are independent then observing x is of no use. We will instead seek to uncover their relationship without any a priori assumptions (non-parametrically) in order to use knowledge of x to inform the choice of control. In Section 2.3 we consider conditional distribution estimators refitted out of existing machine learning methodologies that give rise to estimators of the form (2.4), including k -nearest neighbors methods, kernel interpolation, decision tree (recursive partitioning) methods, and ensemble methods such as random forests. In Section 4 we consider an alternative approach more akin to regularized regression.

Using estimators of the form (2.4), estimating (2.1) yields

$$\min_{z \in \mathcal{Z}} \left\{ \hat{C}_n(z|x) := \sum_{i=1}^n w_n^i(x) c(z; y^i) \right\} \quad (2.5)$$

and estimating (2.3) yields

$$\min_{z \in \mathcal{Z}, \beta \in \mathbb{R}} \left\{ \hat{R}_{\alpha,n}(z, \beta|x) := \beta + \frac{1}{\alpha} \sum_{i=1}^n w_n^i(x) (c(z; Y) - \beta)_+ \right\}. \quad (2.6)$$

Two questions naturally arise: (a) whether (2.5) and (2.6) are efficiently solvable and (b) whether they converge in some sense to (2.1) and (2.3) that they estimate.

2.1 Tractability

Problem (2.5) is similar in complexity to the standard SAA approach (see previous references) and (2.6) is similar to the sample-based approach studied in [34]. For completeness we develop sufficient conditions for either to be solved in polynomial time using the ellipsoid algorithm [20].

Theorem 2.1. *Suppose \mathcal{Z} is a closed convex set and let a separation oracle for it be given.¹ Suppose also that $c(z; y)$ is convex in z for every fixed y and let oracles be given for evaluation and subgradient in z . Then for any fixed x we can find an ϵ -optimal solution to either (2.5) or (2.6) in*

¹E.g., a polyhedron has a trivial separation algorithm.

time and oracle calls polynomial in $n_0, d, \log(1/\epsilon)$ where $n_0 = \sum_{i=1}^n \mathbb{I}[w_n^i(x) > 0] \leq n$ is the effective sample size.

Proof. Let $I = \{i : w_n^i(x) > 0\}$, $w = (w_n^i(x))_{i \in I}$. Rewrite (2.5) as $\min w^T \theta$ over $(z, \theta) \in \mathbb{R}^{d \times n_0}$ subject to $z \in \mathcal{Z}$ and $\theta_i \geq c(z; y^i) \forall i \in I$. Weak optimization of a linear objective over a closed convex body is reducible to weak separation via the ellipsoid algorithm (see [20]). A weak separation oracle for \mathcal{Z} is assumed given. To separate over the i^{th} cost constraint at fixed z' , θ'_i call the evaluation oracle to check violation and if violated call the subgradient oracle to get $s \in \partial_z c(z'; y^i)$ with $\|s\|_\infty \leq 1$ and produce the separating hyperplane $\theta_i \geq c(z'; y^i) + s^T(z - z')$. For (2.6), rewrite the objective as $\min (\beta + w^T \theta / \alpha)$ and change the cost constraints to $\theta_i \geq c(z; y^i) - \beta$, $\theta_i \geq 0 \forall i \in I$. Separation is nearly the same. \square

2.2 Convergence

In terms of convergence there are two concerns: convergence of the optimal value and convergence of the optimal control. Let us write these desired conditions explicitly.

Condition 2.1 (Convergence of value). Almost surely (a.s.) for μ_X -almost-everywhere $x \in \mathcal{X}$ (μ_X -a.e. x),²

$$\begin{aligned} \min_{z \in \mathcal{Z}} \hat{C}_n(z|x) &\rightarrow \min_{z \in \mathcal{Z}} C(z|x) \\ \min_{z \in \mathcal{Z}, \beta \in \mathbb{R}} \hat{R}_{\alpha,n}(z, \beta|x) &\rightarrow \min_{z \in \mathcal{Z}, \beta \in \mathbb{R}} R_\alpha(z, \beta|x). \end{aligned}$$

Condition 2.2 (Convergence of control). A.s. for μ_X -a.e. x ,

$$\begin{aligned} L(\{z_n\}) &\subset \arg \min_{z \in \mathcal{Z}} C(z|x) \\ \text{for all sequences } z_n &\in \arg \min_{z \in \mathcal{Z}} \hat{C}_n(z|x) \\ L(\{(z_n, \beta_n)\}) &\subset \arg \min_{z \in \mathcal{Z}, \beta \in \mathbb{R}} R_\alpha(z, \beta|x) \\ \text{for all sequences } (z_n, \beta_n) &\in \arg \min_{z \in \mathcal{Z}, \beta \in \mathbb{R}} \hat{R}_{\alpha,n}(z, \beta|x) \end{aligned}$$

where L denotes all limit points (a.k.a. accumulation points).

For us, convergence will depend on the consistency of our estimator (2.4) and on the continuity of the cost function. We pose these two conditions below.

Condition 2.3 (Point-wise consistency). For any fixed selection of a single control $z \in \mathcal{Z}$ ($\beta \in \mathbb{R}$), a.s. for μ_X -a.e. x ,

$$\begin{aligned} \int c(z; y) d\hat{\mu}_{Y|x,n}(y) &\rightarrow \int c(z; y) d\mu_{Y|x}(y) \\ \int r_\alpha(c(z; y), \beta) d\hat{\mu}_{Y|x,n}(y) &\rightarrow \int r_\alpha(c(z; y), \beta) d\mu_{Y|x}(y). \end{aligned}$$

Condition 2.4 (Continuity). $c(z; y)$ is equicontinuous in z : for any $z \in \mathcal{Z}$ and $\epsilon > 0$ there exists $\delta > 0$ such that $|c(z; y) - c(z'; y)| \leq \epsilon$ for all z' with $\|z - z'\| \leq \delta$ and $y \in \mathcal{Y}$.

Remark 2.1. A sufficient condition for equicontinuity is that a family of Lipschitz continuous (or, differentiable) functions have bounded Lipschitz constants (or, derivatives). In particular, this is true of the newsvendor and portfolio costs.

Theorem 2.2. *Suppose Conditions 2.3 and 2.4 hold. Then Conditions 2.1 and 2.2 hold.*

²Note that by Fubini's theorem, (((*) holds a.s.) for μ_X -a.e. x) is the same as (((*) holds for μ_X -a.e. x) a.s.)

Proof. By Condition 2.4, $z \mapsto \int c(z; y) d\nu(y)$ is continuous (hence lower-semicontinuous) for any finite measure ν . Let $\{z_i\}$ be any countable dense subset of \mathbb{R}^d (e.g. \mathbb{Q}^d). By Condition 2.3 and since the intersection of countably-many almost sure events is almost sure, we have that

$$\tilde{\xi}_n := \sup_{i \in \mathbb{N}} |\widehat{C}_n(z_i|x) - C(z_i|x)| \rightarrow 0$$

a.s. for μ_X -a.e. x . Define

$$\begin{aligned} \xi_n &:= \sup_{z \in \mathcal{Z}} |\widehat{C}_n(z|x) - C(z|x)| \\ &\leq \tilde{\xi}_n + 2 \sup_{z \in \mathcal{Z}} \inf_{i \in \mathbb{N}} \sup_{y \in \mathcal{Y}} |c(z; y) - c(z_i; y)|. \end{aligned}$$

Let z and $\epsilon > 0$ be given. Then by Condition 2.4 $\exists \delta > 0$ such that $\sup_{y \in \mathcal{Y}} |c(z; y) - c(z_i; y)| \leq \epsilon \forall i : \|z_i - z\| \leq \delta$ and by denseness such i exists. Then $\xi_n = \tilde{\xi}_n \rightarrow 0$ (i.e. uniform convergence) a.s. for μ_X -a.e. x . Since

$$\left| \min_{z \in \mathcal{Z}} \widehat{C}_n(z|x) - \min_{z \in \mathcal{Z}} C(z|x) \right| \leq \xi_n$$

we get Condition 2.1. Now fix a sample path ω and $x \in X$ such that $\xi_n \rightarrow 0$ and suppose there is a sequence $z_n \in \arg \min_{z \in \mathcal{Z}} \widehat{C}_n(z|x)$ and a subsequence $z_{n_k} \rightarrow z_0$ such that $z_0 \notin \arg \min_{z \in \mathcal{Z}} C(z|x)$. Then by continuity

$$\min_{z \in \mathcal{Z}} \widehat{C}_{n_k}(z|x) = \widehat{C}_{n_k}(z_{n_k}|x) \geq C(z_{n_k}|x) - \xi_n \rightarrow C(z_0|x),$$

which, since $C(z_0|x) > \min_{z \in \mathcal{Z}} C(z|x)$, is a contradiction of Condition 2.1 already established, so we get Condition 2.2.

Since $\{\beta + \alpha^{-1}(\cdot - \beta)_+ : \beta \in \mathbb{R}\}$ are uniformly Lipschitz with constant α^{-1} , the composition $\beta + \alpha^{-1}(c(z; y) - \beta)_+$ remains equicontinuous in (z, β) and we can repeat the above for $\widehat{R}_{\alpha, n}(z, \beta|x)$ and $R_\alpha(z, \beta|x)$. \square

The above theorem shows that if we have consistency and continuity then solving our estimated optimization problems (2.5) or (2.6) we will eventually converge to the true optimal policies in both estimated value and recommended control. The optimal policy uses to the utmost the knowledge that $X = x$ and chooses the best z for that scenario. Condition 2.4 depends on our choice of cost function. It remains to be shown that we can come up with estimators that satisfy Condition 2.3.

2.2.1 Sampling Assumptions

The veracity of Condition 2.3 will depend on our choice of estimator and on how we accumulate our sample S_n . We will consider two possible sampling scenarios. One is the traditional assumption that each data point in S_n is independent and identically distributed (IID). This is a strong assumption and is often only a modeling approximation and may fail in practice. The velocity and variety of modern data collection often means that historical observations may not constitute an IID sample. For example, observations taken from an evolving system such as a social network are not IID.

An alternative model of sampling is that the sample S_n is a subsequence taken out of a stationary mixing process.

Definition 2.1. A sequence of random variables V_1, V_2, \dots is called *stationary* if joint distributions of finitely many consecutive variables are invariant to shifting. That is,

$$\mu_{V_t, \dots, V_{t+k}} = \mu_{V_s, \dots, V_{s+k}} \quad \forall s, t \in \mathbb{N}, k \geq 0.$$

In particular, if a sequence is stationary then the variables have identical marginal distributions, but they may not be

independent and the sequence may not be exchangeable. Instead of independence, mixing is the property that if standing at particular point in the sequence we look far ahead enough, the head and the tail look nearly independent, where “nearly” is defined by different metrics for different definitions of mixing.

Definition 2.2. Given a stationary sequence $\{V_t\}_{t \in \mathbb{N}}$, denote by $\mathcal{A}^t = \sigma(V_1, \dots, V_t)$ the sigma-algebra generated by the first t variables and by $\mathcal{A}_t = \sigma(V_t, V_{t+1}, \dots)$ the sigma-algebra generated by the subsequence starting at t . Define the *mixing coefficients at lag k*

$$\alpha(k) = \sup_{t \in \mathbb{N}, A \in \mathcal{A}^t, B \in \mathcal{A}_{t+k}} |\mu(A \cap B) - \mu(A)\mu(B)|$$

$$\beta(k) = \sup_{t \in \mathbb{N}} \left\| \mu\{V_s\}_{s \leq t} \otimes \mu\{V_s\}_{s \geq t+k} - \mu\{V_s\}_{s \leq t} \nu_{s \geq t+k} \right\|_{\text{TV}}$$

$$\rho(k) = \sup_{t \in \mathbb{N}, Q \in L_2(\mathcal{A}^t), R \in L_2(\mathcal{A}_{t+k})} |\text{Corr}(Q, R)|$$

where $L_2(\mathcal{A})$ is the set of \mathcal{A} -measurable square-integrable real-valued random variables.

$\{V_t\}$ is said to be α -mixing if $\alpha(k) \xrightarrow{k \rightarrow \infty} 0$, β -mixing if $\beta(k) \xrightarrow{k \rightarrow \infty} 0$, and ρ -mixing if $\rho(k) \xrightarrow{k \rightarrow \infty} 0$.

Remark 2.2. Notice that an IID sequence has $\alpha(k) = \beta(k) = \rho(k) = 0$. In [8] it is established that $2\alpha(k) \leq \beta(k)$ and $4\alpha(k) \leq \rho(k)$ so that either β - or ρ -mixing implies α -mixing.

Many processes satisfy mixing conditions under mild assumptions: auto-regressive moving-average (ARMA) processes (see [31]), generalized autoregressive conditional heteroskedasticity (GARCH) processes (see [12]), and certain Markov chains. For a thorough discussion and more examples see [15] and [9]. Mixing rates are often given explicitly by model parameters but they can also be estimated from data (see [29]). Sampling from such processes models many real-life sampling situations where observations are taken from an evolving system such as, for example, the stock market, inter-dependent product demands, or a doubly stochastic arrival processes in a social network.

We will often use the following result.

Lemma 2.1. If $\{(x^i, y^i)\}_{i \in \mathbb{N}}$ is stationary and $f : \mathbb{R}^{m_Y} \rightarrow \mathbb{R}$ is measurable then $\{(x^i, f(y^i))\}_{i \in \mathbb{N}}$ is also stationary and has mixing coefficients no larger than those of $\{(x^i, y^i)\}_{i \in \mathbb{N}}$.

Proof. This is simply because a transform can only make the generated sigma-algebra coarser. For a single time point, $\{Y^{-1}(f^{-1}(B)) : B \in \mathcal{B}(\mathbb{R})\} \subset \{Y^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^{m_Y})\}$. Here the transform is applied independently across time. \square

2.3 Estimators

We now review various estimators of the form (2.4) fitted out of existing popular learning machines.

2.3.1 Nearest Neighbor Estimators

In k nearest neighbor (k NN) regression, $\mathbb{E}[Y|X = x]$ is estimated by the average of y^i 's associated with the x^i 's that are among the k nearest neighbors of x in Euclidean distance. Similarly, in k NN classification, we take the mode (most common value) of the y^i 's associated with neighbors. The idea is to replace the missing observations of Y 's for this particular value of $X = x$ that we have not seen before with observations of Y associated with similar values of X .

Reverse engineering these two procedures, we notice that this is exactly the same as computing expectations or modes

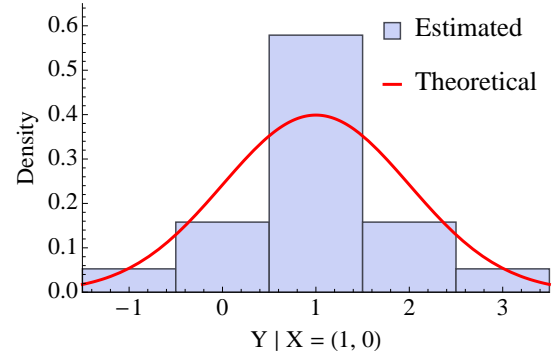
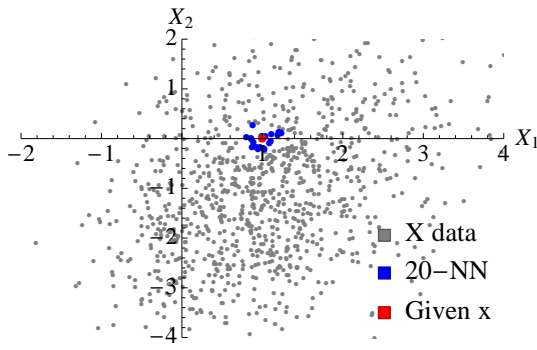


Figure 1: Estimating the conditional distribution using the k NN estimator (2.7) for a simple example and $k = 20$. The unknown underlying structure is $Y = \|X\| + \mathcal{N}(0, 1)$. On the left, the dataset of X values observed, the given $x = (0, 1)$, and its 20 nearest neighbors. On the right, the resulting distribution estimate compared with the true distribution.

over the uniform distribution on these k neighbors. This suggests a conditional distribution estimator based on this procedure of the form of (2.4) with the weights

$$w_n^i(x) = \begin{cases} 1/k & \text{if } x^i \text{ is a } k\text{NN of } x \\ 0 & \text{otherwise} \end{cases} \quad (2.7)$$

Ties are broken either randomly or by a low-index-first rule. An illustration of resulting conditional estimates is given in Figure 1. Incredibly, the resulting estimator is universally consistent (i.e. no assumptions on μ or higher moments of cost) given that we choose k to grow with n but not too fast.

Theorem 2.3. *Suppose S_n is generated by IID sampling. Let $w_n^i(x)$ be as in (2.7) with $k = \min\{\lceil Cn^\delta \rceil, n-1\}$ for some $C > 0$, $0 < \delta < 1$. Then Condition 2.3 holds.*

Proof. This follows directly from Theorem 5 of [41] along with integrability of each $c(z; y)$ for each $z \in \mathcal{Z}$. \square

Finding the k NNs of x without pre-computation can clearly be done in $O(dn)$ time. Data-structures that speed up the process at query time at the cost of pre-computation have been developed (see e.g. [4]) and there are also approximate schemes that can significantly speed up queries (see e.g. [1]).

A variation of nearest neighbor regression is the radius-weighted k -nearest neighbors where observations in the neighborhood are weighted by a decreasing function f in their distance. Constructing a conditional distribution estimator out of this we get an estimator (2.4) with weights

$$w_n^i(x) \propto \begin{cases} f(\|x^i - x\|) & \text{if } x^i \text{ is a } k\text{NN of } x \\ 0 & \text{otherwise} \end{cases}$$

with appropriate normalization to sum to one. Other variations include radius-weighting all sample points and uniformly weighting all sample points within a certain radius. These latter two are subsumed by kernel estimators which we study next.

2.3.2 Kernel Estimators

In non-parametric regression, Nadaraya-Watson (NW) kernel regression [32, 42] estimates $\mathbb{E}[Y|X = x]$ by

$$\frac{\sum_{i=1}^n K((x^i - x)/h) y^i}{\sum_{i=1}^n K((x^i - x)/h)}$$

where $K : \mathbb{R}^d \rightarrow \mathbb{R}$ is a non-negative kernel ($\int K = 1$, $K \geq 0$, and K is unitarily invariant, i.e. it is a symmetric multivariate density) and $h > 0$ is called the bandwidth.

NW kernel regression is based on the conditional distribution that arises from the Parzen-window density estimates of μ and μ_X (i.e., their ratio). In particular, this conditional distribution estimate has the form (2.4) with weights

$$w_n^i(x) = \frac{K((x^i - x)/h)}{\sum_{j=1}^n K((x^j - x)/h)}. \quad (2.8)$$

Some popular kernels are:

1. Naïve: $K(x) = \frac{\Gamma(d/2+1)}{\pi^{d/2}} \mathbb{I}[\|x\| \leq 1]$.
2. Epanechnikov: $K(x) = \frac{\Gamma(d/2+2)}{\pi^{d/2}} (1 - \|x\|^2)_+$.
3. m -weight ($m \in \mathbb{N}$): $K(x) = \frac{\Gamma(d/2+m+1)}{\pi^{d/2} m!} (1 - \|x\|^2)_+^m$.
4. Gaussian: $K(x) = \frac{1}{(2\pi)^{d/2}} \exp(-\|x\|^2/2)$.

Note that the Naïve kernel with bandwidth h corresponds directly to uniformly weighting all neighbors of x that are within a radius of h . The Naïve kernel also results in consistent estimators under certain mild assumptions on μ_X .

Theorem 2.4. *Suppose S_n is generated by IID sampling and that μ_X is a countable mixture of discrete and absolutely continuous measures. Let $w_n^i(x)$ be as in (2.8) with K being the naïve kernel and with $h = Cn^{-\delta}$ for some $C > 0$, $0 < \delta < 1/m_X$. Then Condition 2.3 holds.*

Proof. This follows directly from Theorem 1 of [26] along with integrability of each $c(z; y)$ for each $z \in \mathcal{Z}$. \square

Under stronger integrability conditions, consistency also holds for the other kernels and under the relaxed sampling assumption of mixing and without restrictions on μ_X .

Theorem 2.5. *Let $\phi(t) = (|t| \log |t|)_+$ and suppose $\phi(c(z; y))$ is μ_Y -integrable for each z . Let $w_n^i(x)$ be as in (2.8) with K being any of kernels (1)-(4) and $h = Cn^{-\delta}$ for $C, \delta > 0$. If*

1. S_n is generated by IID sampling and $\delta < 1/m_X$, or
2. S_n comes from a ρ -mixing process with $\rho(k) = O(k^{-\gamma})$ ($\gamma > 0$) and $\delta < 2\gamma/(m_X + 2m_X\gamma)$, or
3. S_n comes from an α -mixing process with $\alpha(k) = O(k^{-\gamma})$ ($\gamma > 1$) and $\delta < 2(\gamma - 1)/(3m_X + 2m_X\gamma)$

then Condition 2.3 holds.

Proof. The integrability conditions on $c(z; y)$ imply the same for $r_\alpha(c(z; y), \beta)$ for every z, β . Then Condition 2.3 follows directly from Theorem 3 of [41] along with Lemma 2.1. \square

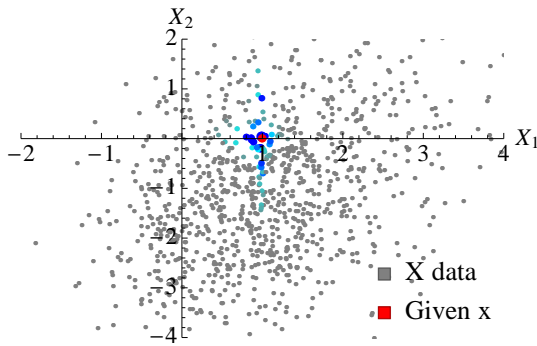


Figure 2: The analogue to Figure 1 for the random forest estimator (2.11) using 100 trees. Bluer colors denote higher weight.

We can avoid the stronger integrability conditions in the above by modifying (2.8) to a semi-recursive estimate where the bandwidth changes with i (and remains fixed for that i):

$$w_n^i(x) = \frac{K((x^i - x)/h_i)}{\sum_{j=1}^n K((x^j - x)/h_j)}. \quad (2.9)$$

Theorem 2.6. Suppose S_n comes from a ρ -mixing process with $\sum_{k=1}^{\infty} \rho(k) < \infty$ (or IID). Let $w_n^i(x)$ be as in (2.9) with K being the naïve kernel and with $h_i = Ci^{-\delta}$ for some $C > 0$, $0 < \delta < 1/(2m_X)$. Then Condition 2.3 holds.

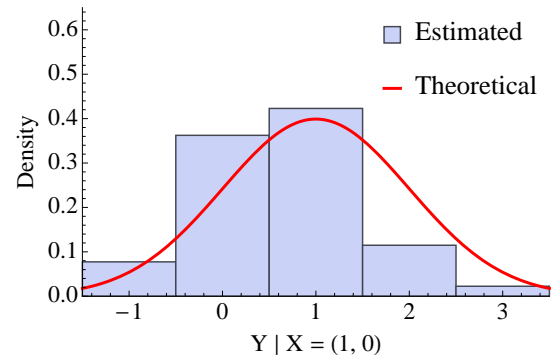
Proof. This follows directly from Theorem 4 of [41] along with Lemma 2.1 and the integrability of each $c(z; y)$. \square

There also exist modifications where observations are weighted according to distance to other observations [33, 18].

2.3.3 Tree Estimators

The above estimators have all been so-called *lazy*, that is, computed on the fly when a new x arrives. This is to be compared to *eager* estimators that compile the data into a structure or description used for future queries. An example of this in classification and regression is classification and regression trees (CART) [10], which recursively split the sample into regions in the space \mathcal{X} so to gain reduction in “impurity” of the response variable within each region. There are different definitions of “impurity,” including Gini and entropy, and different heuristics to choose the best split, different combinations resulting in different algorithms. Multivariate impurity measures are usually the component-wise average of univariate impurities. The value of $\mathbb{E}[Y|X = x]$ (or, the predicted class) is then estimated as the average (or, the mode) of y^i ’s associated with the x^i ’s that reside in the same region as x . The recursive splitting is most often represented as a tree with each non-leaf node representing an intermediate region in the algorithm. As splits are usually restricted to axis-aligned half-spaces, the tree can be represented as subsequent inquiries about whether a particular component of the vector x is larger or smaller than a value. For a thorough review of tree-based methods and their computation see §9.2 of [39].

Regardless of the particular method chosen, the final partition can generally be represented as a rule identifying points in \mathcal{X} with the disjoint regions: $\mathcal{R} : \mathcal{X} \rightarrow \{1, \dots, r\}$. The partition is then the disjoint union $\mathcal{R}^{-1}(1) \sqcup \dots \sqcup \mathcal{R}^{-1}(r) = \mathcal{X}$. The tree regression and classification estimates correspond directly to taking averages or modes over a conditional dis-



tribution estimator of the form (2.4) with weights

$$w_n^i(x) = \frac{\mathbb{I}[\mathcal{R}(x) = \mathcal{R}(x^i)]}{\sum_{j=1}^n \mathbb{I}[\mathcal{R}(x) = \mathcal{R}(x^j)]}. \quad (2.10)$$

Notice that the weights (2.10) are piecewise constant over the partitions and therefore the recommended optimal decision from (2.5) or (2.6) is also piecewise constant. Therefore, solving r optimization problems after the recursive partitioning process, the resulting prescriptive rule can be fully compiled into a decision tree, where the decisions are truly actual decisions. Thus, it retains CART’s interpretability quality. The method can also remain semi-lazy, solving the optimization problem as new x arrive, perhaps with memoization. Apart from being interpretable, tree-based methods are also known to be useful in learning complex interactions and to perform well with large datasets.

2.3.4 Random Forests and Other Ensembles

A random forest is an ensemble of trees each trained on a random subset of components of X . This makes them more uncorrelated and therefore their average have lower variance. Random forests are one of the most popular and flexible tools of machine learning. For a thorough review of random forests and their computation see §15 of [39].

The outcome of a random forest algorithm are multiple partition rules \mathcal{R}_t $t = 1, \dots, T$, one for each tree in the forest. In random forest regression, the regression estimator is the average of the estimate from each tree. By linearity, this is the same as taking averages over a conditional distribution estimator of the form (2.4) with weights

$$w_n^i(x) = \frac{1}{T} \sum_{t=1}^T \frac{\mathbb{I}[\mathcal{R}_t(x) = \mathcal{R}_t(x^i)]}{\sum_{j=1}^n \mathbb{I}[\mathcal{R}_t(x) = \mathcal{R}_t(x^j)]}. \quad (2.11)$$

Other ensemble methods are also extremely popular such as boosting and general bagging. The effect of either boosting or bagging a collection of regressors will generally correspond to taking a convex combination of the corresponding conditional distribution estimators. That is, given a collection of conditional distribution estimators of the form (2.4) with weights $w_{n,t}^i(x)$ for $t = 1, \dots, T$ and $\lambda \in \mathbb{R}^T$, $\lambda \geq 0$, $\sum_{t=1}^T \lambda_t = 1$, the weights

$$w_n^i(x) = \sum_{t=1}^T \lambda_t w_{n,t}^i(x)$$

also form a valid conditional distribution estimator. How to choose λ depends on the ensemble learning method chosen, such as AdaBoost [16] (see also §10 of [39]) or LASSO post-

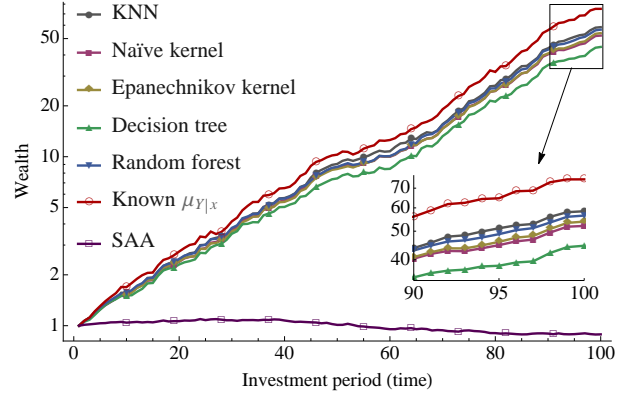
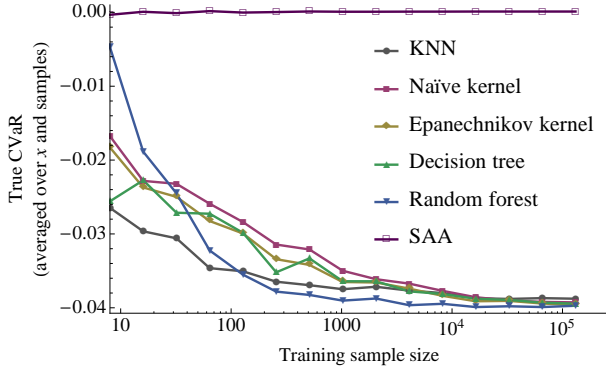


Figure 3: Portfolios minimizing estimated CVaR. On the left, the resulting CVaR by sample size, averaged over samples and new observations x . On the right, accumulated wealth for a single sample path of an evolving market with each method trained on the first 1,000 non-IID points. Note the logarithmic horizontal axis on the left and vertical axis on the right.

processing [17] (see also §16.3 of [39]). These also require an appropriate measure of loss; we discuss this in the following.

2.4 Cross-Validation

There are various tools for comparing different methods, selecting features, and tuning parameters in various learning machines for regression and classification, including those use here to inspire conditional distributional estimators. However, likely the most general and popular approach is cross-validation. Cross-validation partitions the training data and iteratively takes one part out, trains each variation of parameters and features on what is left, and measures its success on the part taken out. The variation chosen is the one that does best overall, but there many ways to measure “best.” In classification f -score, accuracy, and marginal accuracy are all used and in regression squared error is mostly used. It is not always clear what is the appropriate measure because it is also not clear what are the costs of various errors and error magnitudes. In the case studied here actual costs (or, benefits) are the direct object of study and it is clear that if one seeks a prescription that will perform well in terms of costs of decision then the appropriate measure is the true cost on the part of the sample taken out of the prescription trained on the part not taken out. Specifically, consider a partition into training data and validation data:

$$(x^1, y^1), \dots, (x^{n_T}, y^{n_T}) \text{ and } (\hat{x}^1, \hat{y}^1), \dots, (\hat{x}^{n_V}, \hat{y}^{n_V}).$$

Let $w_{n_T}^i(x)$ be the trained conditional distribution estimator and, for the case of expectation optimization, let

$$\hat{z}^i \in \arg \min_{z \in \mathcal{Z}} \hat{C}_{n_T}(z | \hat{x}^i), \quad \hat{v}^i = c(\hat{z}^i; \hat{y}^i).$$

Then we measure the validation loss as the average of \hat{v}^i . This, in turn, we can average over cross-validation folds. Note that choosing a prescription in this way is effectively the same as if we had used discrepancy to the true optimum $\hat{u}^i = |c(\hat{z}^i; \hat{y}^i) - \min_{z \in \mathcal{Z}} c(z; \hat{y}^i)|$ because by definition of the optimum (always smaller), $\hat{u}^i = c(\hat{z}^i; \hat{y}^i) - \min_{z \in \mathcal{Z}} c(z; \hat{y}^i)$, so that differences of this measure between prescriptions are the same as differences in \hat{v} .

3. CASE STUDY: PORTFOLIO PROBLEM

Let us return to the portfolio example from the introduction. We had d securities, Y_i was the future return of security i in percents ($\mathcal{Y} = \mathbb{R}^d$), and z_i was the fraction invested in

security i ($\mathcal{Z} = \{z \in \mathbb{R}^d : z \geq 0, \sum_i z_i = 1\}$). Thus, the portfolio z will have returns $z^T Y$ ($c(z; y) = -z^T y$).

Let us consider a particular simple instance. Let $d = 8$ and suppose returns are generated according to a factor model (e.g. as in the Capital Asset Pricing Model) $Y_i = A_i^T X + W_i$, where $X \in \mathbb{R}^3$ represents common market factors,

$$A_i = \left((-1)^{\lceil i/4 \rceil} \%, (-1)^{\lceil i/2 \rceil} \%, (-1)^i \% \right) \quad i = 1, \dots, 8$$

is the unique dependence of the i^{th} security on these common factors, and $W_i \sim \mathcal{N}(0\%, \Sigma)$ is an idiosyncratic contribution with covariance $\Sigma_{ij} = (\mathbb{I}[i = j] \frac{8}{7} - (-1)^{i+j} \frac{1}{7}) 0.05 (\%^2)$. In accordance with the fact that the stock market is an evolving system, we consider factors evolving over time as a 3-dimensionanl ARMA(2,2) process:

$$X(t) - \Phi_1 X(t-1) - \Phi_2 X(t-2) = U(t) + \Theta_1 U(t-1) + \Theta_2 U(t-2)$$

where $U \sim \mathcal{N}(0, \Sigma_U)$ are Gaussian innovations. We give the values of the 3×3 matrices $\Phi_1, \Phi_2, \Theta_1, \Theta_2, \Sigma_U$ in the appendix. Marginally, all of the returns are normally distributed with mean 0% and standard deviations 2.5~3%.

We consider a situation where one can observe the factors X but is not aware of the dependence structure and is to completely non-parametrically learn this from non-IID observations in order to construct a portfolio with minimal conditional value at risk at level $\alpha = 15\%$. We consider each of the methods presented in the Section 2.3, as well as the omniscient optimum (2.3) and the marginal SAA that ignores the factors. In the left panel of Figure 3 we report the average performance of each of these. In right panel we plot the wealth trajectory of each of these methods (trained on a 1,000 non-IID samples) in a dynamic trading scenario where one starts with unit wealth and in each period invests one’s whole wealth in the method’s recommended portfolio.

4. RULE-BASED PRESCRIPTIONS

In the previous sections we considered a decision rule that finds an approximately optimal decision z separately for each x . Another approach would be to develop upfront a decision rule that constructs a more explicit mapping from observations to decisions. The set up is as follows. We are interested in choosing a decision rule $\zeta : \mathcal{X} \rightarrow \mathcal{Z}$ out of a family of possible ones \mathcal{F} so as either to minimize marginal empirical

expected costs,

$$\min_{\zeta \in \mathcal{F}} \left\{ \widehat{C}_n(\zeta) := \frac{1}{n} \sum_{i=1}^n c(\zeta(x^i); y^i) \right\}, \quad (4.1)$$

or to minimize marginal empirical CVaR at level α ,

$$\min_{\zeta \in \mathcal{F}, \beta \in \mathbb{R}} \left\{ \widehat{R}_{\alpha, n}(\zeta, \beta) := \beta + \frac{1}{\alpha n} \sum_{i=1}^n \left(c(\zeta(x^i); y^i) - \beta \right)_+ \right\}. \quad (4.2)$$

After training ζ as above, when we observe x we will make the decision $z = \zeta(x)$. The hope is that the above chooses a rule ζ that has small real-world overall expected costs $C(\zeta) = \mathbb{E}[c(\zeta(X); Y)]$ or risk $\text{CVaR}_\alpha(c(\zeta(X); Y))$. In the particular case of the univariate single-item newsvendor problem ($\mathcal{Z} = \mathbb{R}$), this is similar to the approach taken in [35]. We treat a more general problem with general cost functions (or CVaR) and multivariate decisions that are possibly constrained and with more general functional families \mathcal{F} .

Examples of \mathcal{F} include:

1. Linear combinations of features. Let $f : \mathcal{X} \rightarrow \mathbb{R}^T$ be a feature mapping and consider

$$\mathcal{F} = \left\{ \zeta(x) = Wf(x) : W \in \mathbb{R}^{d \times T}, \|W\| \leq R \right\} \quad (4.3)$$

for some choice of norm such as, for example, a row-wise p, p' -norm

$$\|W\| = \left\| \left(\|W_1\|_p / \gamma_1, \dots, \|W_d\|_p / \gamma_d \right) \right\|_{p'},$$

or a row-interacting Schatten p -norm:

$$\|W\|_p = \sqrt[p]{\sum_i \tau_i^p} \text{ where } \tau_i \text{ are } W\text{'s singular values.}$$

For example, $p = 1$ corresponds to the rank-sparsity-inducing nuclear norm.

2. Product of reproducing kernel Hilbert spaces (RKHS).

$$\mathcal{F} = \{ \zeta(x) = (\zeta_1(x), \dots, \zeta_d(x)) : \zeta_i \in \mathcal{H}_i, \|\zeta_i\| \leq R_i \} \quad (4.4)$$

for some choice of RKHSs \mathcal{H}_k . This generalizes the above in the case $p = 2, p' = \infty$ to infinite dimensions. A Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is an inner product space that is a separable Banach space with the norm defined by $\|f\|^2 = \langle f, f \rangle$. An RKHS is a Hilbert space for which evaluations are continuous. The Riesz representation theorem then immediately yields that for each $x \in \mathcal{X}$ there is $\mathcal{K}(x, \cdot) \in \mathcal{H}$ such that $\langle \mathcal{K}(x, \cdot), h(\cdot) \rangle = h(x)$ for every $h \in \mathcal{H}$. The symmetric map $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called the reproducing kernel, the name motivated by the fact that $\mathcal{H} = \text{closure}(\text{span}\{\mathcal{K}(x, \cdot)\}_x)$. See [5] for more details on RKHS. Examples include:

- (a) The polynomial kernel $\mathcal{K}_s(x, x') = (1 + x^T x' / s)^s$ is an example of the previous, spanning all monomials of degree up to s .
- (b) Any kernel $\mathcal{K}(x, x') = \sum_{i=0}^{\infty} a_i (x^T x')^i$ with $a_i \geq 0$ (subject to convergence) such as the previous or the exponential kernel $\mathcal{K}(x, x') = e^{x^T x'}$, which can be seen as its infinite-dimensional limit.
- (c) The Gaussian kernel $\mathcal{K}_s(x, x') = e^{-s^2 \|x - x'\|^2}$ (the corresponding space \mathcal{H} is studied in [38]).

In both cases the restrictions on the norm are equivalent to incorporating an appropriately-weighted regularization term

into the objectives of (4.1) and (4.2).³ This bears similarity to regularized multiple regression. Setting $\mathcal{Z} = \mathcal{Y}$ and $c(z; y) = (z - y)^2$ does not quite fit our understanding of $c(z; y)$ as a cost of making a decision but it recovers least-squares multiple regression in the case of expectation minimization (4.1). Below we develop sufficient conditions for either (4.1) or (4.2) with these examples be solved in polynomial time using the ellipsoid algorithm [20] when $\mathcal{Z} = \mathbb{R}^d$.

Theorem 4.1. *Suppose that $c(z; y)$ is convex in z for every fixed y and let oracles be given for evaluation and subgradient in z . Then for any fixed x we can find an ϵ -optimal solution to (4.1) or (4.2) in time and oracle calls polynomial in $n, d, T, \log(1/\epsilon)$ for (4.3) or in $n, d, \log(1/\epsilon)$ for (4.4).*

Proof. In the case of (4.3), letting $F_{ti} = f_t(x^i)$, we get $\zeta_j(x^i) = W F e_i$ where e_i is the i^{th} unit vector. Also, by computing the norm we have a trivial weak membership algorithm for the norm constraint and hence by Theorems 4.3.2 and 4.4.4 of [20] we also have a weak separation algorithm. Consider the case of (4.4). Since (4.1) and (4.2) only depend on $\zeta \in \mathcal{F}$ through its evaluations at data points $\zeta_j(x^i) = \langle \zeta_j, \mathcal{K}_j(x^i, \cdot) \rangle_{\mathcal{H}_j}$ it is clear that we can restrict to the finite-dimensional subspaces

$$\mathcal{A}_j = \text{span}(\mathcal{K}_j(x^1, \cdot), \dots, \mathcal{K}_j(x^n, \cdot))$$

since adding anything in the orthogonal complement can only inflate magnitudes and leaves evaluations unchanged. Therefore, we may switch to optimizing over the variables $\Gamma \in \mathbb{R}^{d \times n}$ and let $\zeta_j(x) = \sum_{i=1}^n \Gamma_{ji} \mathcal{K}_j(x^i, x)$. Then $\zeta_j(x^i) = \Gamma_j^T K_j e_i$ where $K_{jii'} = \mathcal{K}_j(x^i, x^{i'})$ is the kernel Gram matrix. Also $\|\zeta_j\| = \|\Gamma_j\|_2$ and we have a separation algorithm as before. In either case, by adding affine constraints $z_{ij} = \zeta_j(x^i)$, all that is left is to separate over constraints of the form $\theta_i \geq c(z_i; y^i)$, which was covered in Theorem 2.1. \square

If \mathcal{Z} is a constrained set, it is difficult to use the above to express a general and effective decision rule family that takes values only in \mathcal{Z} . For this purpose we may generally consider composing the above with projections onto \mathcal{Z} . Suppose \mathcal{Z} is a closed convex set and let

$$\Pi_{\mathcal{Z}}(z') = \arg \min_{z \in \mathcal{Z}} \|z - z'\|.$$

Paraphrasing Proposition 2.2.1 from [6],

Theorem 4.2. *For each $z' \in \mathbb{R}^d$, $\Pi_{\mathcal{Z}}(z')$ exists and is a singleton (z exists and is unique). For each $z', z'' \in \mathbb{R}^d$,*

$$\|\Pi_{\mathcal{Z}}(z') - \Pi_{\mathcal{Z}}(z'')\| \leq \|z' - z''\|.$$

Then we will generally consider rules

$$\zeta(x) = \Pi_{\mathcal{Z}}(\zeta'(x)) \text{ s.t. } \zeta' \in \mathcal{F}$$

where \mathcal{F} has any domain, such as the examples above. In this section we will operator under the following assumption:

Assumption 4.1. Either \mathcal{Z} is a closed convex set or the range of each $\zeta \in \mathcal{F}$ is contained in \mathcal{Z} (projection is identity).

We will also abuse notation and define $C(z) = C(\Pi_{\mathcal{Z}}(z))$ whenever $z \notin \mathcal{Z}$ and similarly for $\widehat{C}_n, R_\alpha, \widehat{R}_{\alpha, n}$. Optimization may be more difficult as these may no longer be convex.

Notice that in the case of expectation minimization, this approach coincides with conditional-distribution-based prescriptions when here \mathcal{F} is taken to be the unconstrained

³This can be seen by considering the necessary first-order conditions and seeing that the stationary points coincide.

space of all functions $\mathcal{X} \rightarrow \mathcal{Z}$ and in the other $\hat{\mu}_{Y|x,n}$ is taken to be the empirical conditional distribution estimate. Under this choice, the decision for any x not previously seen ($x \neq x^i \forall i$) is completely under-determined in this approach and is undefined in the other. Therefore we can interpret the two approaches as different ways to smooth out the discrete data. Here we directly restrict the smoothness of the decision function and in the previous we smooth out our estimate of the conditional distribution by employing appropriate estimators fitted out of machine learning methods.

4.1 Out-of-Sample Guarantees

We will characterize out-of-sample guarantees in terms of a multivariate modification of Rademacher complexity.

Definition 4.1. Given a sample $S_n = \{s_1, \dots, s_n\}$, The empirical multivariate Rademacher complexity of a class of functions \mathcal{G} taking values in \mathbb{R}^d is defined as

$$\hat{\mathfrak{R}}_n(\mathcal{G}; S) = \mathbb{E} \left[\frac{2}{n} \sup_{g \in \mathcal{G}} \sum_{i=1}^n \sum_{k=1}^d \sigma_{ik} g_k(s_i) \middle| s_1, \dots, s_n \right]$$

where σ_{ik} are independently equiprobably $+1, -1$. The marginal multivariate Rademacher complexity is defined as

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E} [\hat{\mathfrak{R}}_n(\mathcal{G}; S)]$$

over the sampling distribution of S_n (e.g. if the sample is IID from μ then the sampling distribution is μ^n).

Notice that when $d = 1$ the above definition coincides with the common definition of Rademacher complexity (see [3]). In addition, it is clear that

$$\hat{\mathfrak{R}}_n(\mathcal{G}; S) \leq \sum_{k=1}^d \hat{\mathfrak{R}}_n(\Pi_k \circ \mathcal{G}; S)$$

where Π_k is the projection onto the k^{th} component and the right-hand-side complexities are the common univariate Rademacher complexities. In particular, we have equality if \mathcal{G} is the cartesian product of univariate function classes.

The theorem below is a direct application of common results for univariate complexities that follow easily from McDiarmid's inequality [3] and a mixing variation of those [30].

Theorem 4.3. Suppose c is bounded

$$\sup_{z \in \mathcal{Z}, y \in \mathcal{Y}} c(z; y) \leq \bar{c}.$$

Let $\mathcal{G} = \{(x, y) \mapsto c(\Pi_{\mathcal{Z}}(f(x)); y) : f \in \mathcal{F}\}$. Fix $\delta > 0$. If S_n is generated by IID sampling, let $\delta' = \delta'' = \delta$ and $\nu = n$. If S_n comes from a β -mixing process, fix some t, ν such that $2t\nu = n$, let $\delta' = \delta/2 - (\nu-1)\beta(t)$ and $\delta'' = \delta/2 - 2(\nu-1)\beta(t)$. Then (only for $\delta' > 0$ or $\delta'' > 0$ where they appear),

$$\begin{aligned} \bullet \quad C(\zeta) &\leq \hat{C}_n(\zeta) + \bar{c}\sqrt{\log(1/\delta')/2\nu} + \mathfrak{R}_\nu(\mathcal{G}) \\ \forall \zeta \in \mathcal{F}, \text{ with probability at least } 1 - \delta. \end{aligned} \quad (4.5)$$

$$\begin{aligned} \bullet \quad C(\zeta) &\leq \hat{C}_n(\zeta) + 3\bar{c}\sqrt{\log(2/\delta'')/2\nu} + \hat{\mathfrak{R}}_\nu(\mathcal{G}; x, y) \\ \forall \zeta \in \mathcal{F}, \text{ with probability at least } 1 - \delta. \end{aligned} \quad (4.6)$$

$$\begin{aligned} \bullet \quad R_\alpha(\zeta, \beta) &\leq \hat{R}_{\alpha,n}(\zeta, \beta) + \frac{\bar{c}}{\alpha}\sqrt{\log(1/\delta')/2\nu} + \mathfrak{R}_\nu(\mathcal{G}) \\ \forall \zeta \in \mathcal{F}, \beta \in \mathbb{R}, \text{ with probability at least } 1 - \delta. \end{aligned} \quad (4.7)$$

$$\begin{aligned} \bullet \quad R_\alpha(\zeta, \beta) &\leq \hat{R}_{\alpha,n}(\zeta, \beta) + \frac{3\bar{c}}{\alpha}\sqrt{\log(2/\delta'')/2\nu} + \hat{\mathfrak{R}}_\nu(\mathcal{G}; x, y) \\ \forall \zeta \in \mathcal{F}, \beta \in \mathbb{R}, \text{ with probability at least } 1 - \delta. \end{aligned} \quad (4.8)$$

Proof. In the case of IID sampling, the first two results are classical results for univariate Rademacher complexity (see e.g. [3]) and the latter two results follow by using the univariate comparison lemma (Theorem 4.12 of [27]) using the 1-Lipschitz univariate transformation $c \mapsto (c)_+$ and applying the standard result to the marginal expectation inside the variational formulation of CVaR of in (2.2). The modifications for β -mixing are due to Theorems 1 and 2 of [30]. \square

However, how to compute $\hat{\mathfrak{R}}_n(\mathcal{G}; x, y)$ or what is its relationship to \mathcal{F} are both concerns. We begin by addressing the second question. In the following we adapt Theorem 4.12 of [27] to our multivariate case.

Lemma 4.1. Suppose that c is L -Lipschitz uniformly over y with respect to ∞ -norm:

$$\sup_{z \neq z' \in \mathcal{Z}, y \in \mathcal{Y}} \frac{c(z; y) - c(z'; y)}{\max_{k=1, \dots, d} |z_k - z'_k|} \leq L < \infty.$$

Then we have that $\hat{\mathfrak{R}}_n(\mathcal{G}; x, y) \leq L\hat{\mathfrak{R}}_n(\mathcal{F}; x)$ for \mathcal{G} as in Theorem 4.3 and therefore also that $\mathfrak{R}_n(\mathcal{G}) \leq L\mathfrak{R}_n(\mathcal{F})$. (Notice that one is a univariate complexity and one multivariate and that the complexity of \mathcal{F} involves only the sampling of x .)

Proof. Write $\phi_i(z) = c(\Pi_{\mathcal{Z}}(z); y_i)/L$. Then by Lipschitz assumption and by part 2 of Theorem 4.2, for each i , ϕ_i is 1-Lipchitz. We now would like to show the inequality in

$$\begin{aligned} \hat{\mathfrak{R}}_n(\mathcal{G}; x, y) &= L\mathbb{E} \left[\frac{2}{n} \sup_{\zeta \in \mathcal{F}} \sum_{i=1}^n \sigma_{i0} \phi_i(\zeta(x_i)) \middle| x, y \right] \\ &\leq L\mathbb{E} \left[\frac{2}{n} \sup_{\zeta \in \mathcal{F}} \sum_{i=1}^n \sum_{k=1}^d \sigma_{ik} \zeta_k(x_i) \middle| x \right] \\ &= L\hat{\mathfrak{R}}_n(\mathcal{F}; x). \end{aligned}$$

By conditioning and iterating, it suffices to show that for any $T \subset \mathbb{R} \times \mathcal{Z}$ and 1-Lipchitz ϕ ,

$$\mathbb{E} \left[\sup_{t, z \in T} (t + \sigma_0 \phi(z)) \right] \leq \mathbb{E} \left[\sup_{t, z \in T} \left(t + \sum_{k=1}^d \sigma_k z_k \right) \right]. \quad (4.9)$$

The expectation on the left-hand-side is over two values ($\sigma_0 = \pm 1$) so there are two choices of (t, z) , one for each scenario. Let any $(t^{(+1)}, z^{(+1)}), (t^{(-1)}, z^{(-1)}) \in T$ be given. Let k^* and $s^* = \pm 1$ be such that

$$\max_{k=1, \dots, d} |z_k^{(+1)} - z_k^{(-1)}| = s^* (z_{k^*}^{(+1)} - z_{k^*}^{(-1)}).$$

Fix $(\tilde{t}^{(\pm 1)}, \tilde{z}^{(\pm 1)}) = (t^{(\pm s^*)}, z^{(\pm s^*)})$. Then, since these are feasible choices in the inner supremum, choosing $(t, z)(\sigma) = (\tilde{t}^{(\sigma_{k^*})}, \tilde{z}^{(\sigma_{k^*})})$, we see that the right-hand-side of (4.9) has

$$\begin{aligned} \text{RHS (4.9)} &\geq \frac{1}{2} \mathbb{E} \left[\tilde{t}^{(+1)} + \tilde{z}_{k^*}^{(+1)} + \sum_{k \neq k^*} \sigma_k \tilde{z}_k^{(+1)} \right] \\ &\quad + \frac{1}{2} \mathbb{E} \left[\tilde{t}^{(-1)} - \tilde{z}_{k^*}^{(-1)} + \sum_{k \neq k^*} \sigma_k \tilde{z}_k^{(-1)} \right] \\ &= \frac{1}{2} \left(t^{(+1)} + t^{(-1)} + \max_{k=1, \dots, d} |z_k^{(+1)} - z_k^{(-1)}| \right) \\ &\geq \frac{1}{2} \left(t^{(+1)} + \phi(z^{(+1)}) \right) + \frac{1}{2} \left(t^{(-1)} - \phi(z^{(-1)}) \right) \end{aligned}$$

where the last inequality is due to the Lipschitz condition. Since true for any $(t^{(\pm 1)}, z^{(\pm 1)})$ given, taking suprema over the left-hand-side completes the proof. \square

We can therefore bound $\widehat{\mathfrak{R}}_n(\mathcal{G}; x, y)$ if we can do the same for $\widehat{\mathfrak{R}}_n(\mathcal{F}; x)$. In the case of linear families, [23] provides bounds for the univariate components. Applying Theorem 3 of [23] to each component yields the following result:

Lemma 4.2. *Consider \mathcal{F} as in (4.3) with row-wise p, p' norm for $p \in [2, \infty)$ and $p' \in [1, \infty]$. Let q be the conjugate exponent of p ($1/p + 1/q = 1$) and suppose that $\|f(x)\|_q \leq F$ for all $x \in \mathcal{X}$. Then*

$$\mathfrak{R}_n(\mathcal{F}) \leq 2FR\sqrt{\frac{q-1}{n}} \sum_{k=1}^d \gamma_k.$$

The case of a product of RKHSs is similarly treated.

Lemma 4.3. *Consider \mathcal{F} as in (4.4). Then*

$$\begin{aligned} \widehat{\mathfrak{R}}_n(\mathcal{F}; x) &\leq \frac{2}{\sqrt{n}} \sum_{k=1}^d R_k \sqrt{\widehat{\mathbb{E}}_n K_k(x, x)} \\ \mathfrak{R}_n(\mathcal{F}) &\leq \frac{2}{\sqrt{n}} \sum_{k=1}^d R_k \sqrt{\mathbb{E} K_k(x, x)} \end{aligned}$$

where $\widehat{\mathbb{E}}_n$ denotes expectation with respect to the empirical distribution of the sample x_1, \dots, x_n .

Proof. We work component-wise. By Jensen's inequality,

$$\begin{aligned} \widehat{\mathfrak{R}}_n(\{\|h\|_{\mathcal{H}_k} \leq R_k\}; x) &\leq R_k^2 \mathbb{E} \left[\left\| \frac{2}{n} \sum_{i=1}^n \sigma_i \mathcal{K}_k(x_i, \cdot) \right\|_{\mathcal{H}_k}^2 \middle| x \right] \\ &= R_k^2 \frac{4}{n^2} \sum_{i,j=1}^n \mathbb{E} [\sigma_i \sigma_j] \mathcal{K}_k(x_i, x_j) \\ &= R_k^2 \frac{4}{n^2} \sum_{i=1}^n \mathcal{K}_k(x_i, x_i). \end{aligned}$$

The second result follows by applying Jensen's inequality again to pass the expectation over S_n into the square. \square

Finally we address the case of (4.3) with Schatten norm.

Lemma 4.4. *Consider \mathcal{F} as in (4.3) with Schatten p -norm for $p \leq 2$. Then*

$$\begin{aligned} \widehat{\mathfrak{R}}_n(\mathcal{F}; x) &\leq 2R\sqrt{\frac{d}{n}} \sqrt{\widehat{\mathbb{E}}_n \|f(x)\|_2^2} \\ \mathfrak{R}_n(\mathcal{F}) &\leq 2R\sqrt{\frac{d}{n}} \sqrt{\mathbb{E} \|f(x)\|_2^2}. \end{aligned}$$

Proof. Let $F_{ti} = f_t(x_i)$ and let q be p 's conjugate exponent ($1/p + 1/q = 1$) then since $q \geq 2$ and by Jensen's inequality

$$\begin{aligned} \widehat{\mathfrak{R}}_n^2(\mathcal{F}; x) &\leq \frac{4}{n^2} \mathbb{E} \left[\sup_{\|W\|_p \leq R} \text{Trace}(WF\sigma)^2 \middle| x \right] \\ &= \frac{4R^2}{n^2} \mathbb{E} [\|F\sigma\|_q^2 | x] \\ &\leq \frac{4R^2}{n^2} \mathbb{E} [\|F\sigma\|_2^2 | x]. \end{aligned}$$

The first result follows because

$$\begin{aligned} \frac{1}{n} \mathbb{E} [\|F\sigma\|_2^2 | x] &= \frac{1}{n} \sum_{k=1}^d \sum_{t=1}^T \sum_{i,i'=1}^n F_{ti} F_{ti'} \mathbb{E} [\sigma_{ik} \sigma_{i'k}] \\ &= \frac{d}{n} \sum_{i=1}^n \sum_{t=1}^T F_{ti}^2 = d \widehat{\mathbb{E}}_n \|f(x)\|_2^2 \end{aligned}$$

The second result follows by applying Jensen's inequality again to pass the expectation over S_n into the square. \square

The above results can also be combined where applicable. For example, if the decision z logically decomposes into two parts z_1 and z_2 we may wish to have different families for each and consider their product (e.g. using different features and/or having separate simultaneous sparsity constraints).

5. CONCLUSION

We addressed how to transform predictive learning machines into prescriptive mechanisms that find optimal decisions based on predictive observations that foretell what the future may hold and historical data that gives us a glimpse into the possible relationship between these observations and future cost parameters. This is a departure from most of the existing literature on data-driven optimization where predictive observations are usually not employed. This change introduced the difficulty of estimating conditional distributions for given observations we have never before observed. The parallel in prediction is the estimation of conditional expectations given observations not in the training sample. In one approach we considered some of the most popular non-parametric predictive mechanisms that successfully thus generalize data and we refitted conditional-distribution estimators out of them for use in a stochastic-conditional optimization problem. Making mathematical connections between statistical pointwise consistency of regression on a single quantity and the convergence of these optimization problems, we were able to show that for almost all data samples and almost any given observation the decisions and their cost/risk estimates will converge to those of the optimal omniscient policy that, having knowledge of the unknown conditional distributions, uses the observations to their fullest extent in choosing the optimal decision. Incredibly, in many cases this convergence persisted even when data was not IID but rather drawn from mixing processes that model evolving systems. We also considered an alternative approach that is more similar to regularized regression and can be seen as an alternative way to smooth the underdetermined problem when a new observation is not the training sample in order to come up with a well-defined recommended decision. We showed how to develop out-of-sample guarantees in this case that ensure that the real-world costs/risks are similar to the estimated ones. However, this approach is perhaps less universal as it may not converge to the omniscient optimal decision rule without assumptions on its functional form and optimization may be more difficult when the decision-space is constrained. Both of these issues are not problems in the former approach that enjoys non-parametric universal convergence and a simple optimization problem even for constrained decisions. With data permeating every level of life and of business and shown to predict many future quantities of interest, the methods presented herein have the capacity to turn predictions based on this data into optimal decisions that also make full use of all data available.

6. REFERENCES

- [1] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *J. ACM*, 45(6):891–923, 1998.

- [2] S. Asur and B. Huberman. Predicting the future with social media. In *WI-IAT*, 2010.
- [3] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2003.
- [4] J. Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, 1975.
- [5] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, 2004.
- [6] D. Bertsekas, A. Nedić, and A. Ozdaglar. *Convex analysis and optimization*. Athena Scientific, Belmont, 2003.
- [7] J. R. Birge and F. Louveau. *Introduction to stochastic programming*. Springer, 2011.
- [8] R. Bradley. Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*, pages 165–192. Birkhauser, 1986.
- [9] R. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surv.*, 2(107-44):37, 2005.
- [10] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and regression trees*. CRC press, 1984.
- [11] H. A. Carneiro and E. Mylonakis. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin. Infect. Dis.*, 49(10):1557–1564, 2009.
- [12] M. Carrasco and X. Chen. Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18(1):17–39, 2002.
- [13] H. Choi and H. Varian. Predicting the present with google trends. *Econ. Rec.*, 88(s1):2–9, 2012.
- [14] Z. Da, J. Engelberg, and P. Gao. In search of attention. *J. Finance*, 66(5):1461–1499, 2011.
- [15] P. Doukhan. *Mixing: Properties and Examples*. Springer, 1994.
- [16] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Second European Conference on Computational Learning Theory*.
- [17] J. Friedman and B. Popescu. Importance sampled learning ensembles, 2003. Technical report.
- [18] T. Gasser and H.-G. Müller. Kernel estimation of regression functions. pages 23–68, 1979.
- [19] S. Goel, J. Hofman, S. Lahaie, D. Pennock, and D. Watts. Predicting consumer behavior with web search. *PNAS*, 107(41):17486–17490, 2010.
- [20] M. Grotschel, L. Lovasz, and A. Schrijver. *Geometric algorithms and combinatorial optimization*. Springer, New York, 1993.
- [21] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien. How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Syst. J.*, 43(1):64–77, 2004.
- [22] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *SIGKDD*, 2005.
- [23] S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, pages 793–800, 2008.
- [24] N. Kallus. Predicting crowd behavior with big public data. In *WWW*, 2014.
- [25] A. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.*, 12(2):479–502, 2002.
- [26] A. Kozek, J. Leslie, and E. Schuster. On a universal strong law of large numbers for conditional expectations. *Bernoulli*, 4(2):143–165, 1998.
- [27] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.
- [28] A. McAfee and E. Brynjolfsson. Big data: the management revolution. *Harvard Bus. Rev.*, 90(10):60–66, 2012.
- [29] D. McDonald, C. Shalizi, and M. Schervish. Estimating beta-mixing coefficients. In *AISTATS*, pages 516–524, 2011.
- [30] M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-iid processes. In *NIPS*, pages 1097–1104, 2008.
- [31] A. Mokkadem. Mixing properties of arma processes. *Stochastic Process. Appl.*, 29(2):309–315, 1988.
- [32] E. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.
- [33] M. Priestley and M. Chao. Non-parametric function fitting. *J. R. Stat. Soc. Ser. B*, pages 385–392, 1972.
- [34] T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *J. Risk*, 2:21–42, 2000.
- [35] C. Rudin and G.-Y. Vahn. The big data newsvendor: Practical insights from machine learning. 2014.
- [36] A. Shapiro. Monte carlo sampling methods. *Handbooks Oper. Res. Management Sci.*, 10:353–425, 2003.
- [37] A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer, 2005.
- [38] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory*, 52(10):4635–4643, 2006.
- [39] H. Trevor, T. Robert, and J. Friedman. *The Elements of Statistical Learning*, volume 1. Springer, 2001.
- [40] L. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.
- [41] H. Walk. Strong laws of large numbers and nonparametric estimation. In *Recent Developments in Applied Probability and Statistics*, pages 183–214. Springer, 2010.
- [42] G. Watson. Smooth regression analysis. *Sankhyā A*, pages 359–372, 1964.